DOI 10.37882/2223-2966.2026.03.14

ЭМБЕДДИНГИ КАК ОСНОВА АВТОМАТИЧЕСКОГО ИНДЕКСИРОВАНИЯ НАУЧНЫХ ТЕКСТОВ КЛЮЧЕВЫМИ ТЕРМИНАМИ: АЛГОРИТМИЧЕСКИЕ ОГРАНИЧЕНИЯ И СТРУКТУРНО-СЕМАНТИЧЕСКАЯ МОДЕЛЬ

EMBEDDINGS AS THE BASIS FOR AUTOMATIC INDEXING OF SCIENTIFIC TEXTS WITH KEY TERMS: ALGORITHMIC CONSTRAINTS AND A STRUCTURAL AND SEMANTIC MODEL

I. Komarov

Summary. This article discusses the features of embedding in the tasks of automatic indexing of scientific texts with key terms. To analyze the possibilities and limitations of embedding-oriented indexing methods, three groups of factors are identified: equality of vectors, contextual blurriness, loss of structural significance of terms, and a structural and semantic model using a weighted representation of the term and an aggregated document vector is proposed. The results of the study showed that the best quality and reproducibility of indexing is achieved by integrating the semantic proximity of candidates, considering the structure of the scientific text, thereby increasing the consistency of key terms in digital scientific collections.

Keywords: automatic indexing, embeddings, keywords, vector representations, scientific texts, structural and semantic model, digital scientific collections.

Комаров Иван Дмитриевич
Аспирант, Всероссийский институт научной и технической информации Российской академии наук, г. Москва
i.komaroni@ya.ru

Аннотация. В данной статье рассматриваются особенности применения эмбеддингов в задачах автоматического индексирования научных текстов ключевыми терминами. Для анализа возможностей и ограничений эмбеддинг-ориентированных методов индексирования выделены три группы факторов: равноправие векторов, контекстная размытость, утрата структурной значимости терминов, а также предложена структурно-семантическая модель с использованием взвешенного представления терма и агрегированного вектора документа. Результаты исследования показали, что наилучшее качество и воспроизводимость индексирования достигаются при интеграции семантической близости кандидатов с учётом структуры научного текста, посредством чего обеспечивается повышение согласованности ключевых терминов в цифровых научных коллекциях.

Ключевые слова: автоматическое индексирование, эмбеддинги, ключевые термины, векторные представления, научные тексты, структурно-семантическая модель, цифровые научные коллекции.

Введение

Автоматическое индексирование научных текстов в цифровых научных коллекциях относится к задачам системного анализа, поскольку результат определяется согласованием этапов представления текста, отбора признаков и принятия решения о наборе ключевых терминов, пригодном для поиска и аналитической обработки [12]. Эмбеддинги в этом смысле рассматриваются как способ перевода лексических единиц и фрагментов текста в числовую форму, что позволяет работать с семантической близостью терминов в векторном пространстве на больших массивах публикаций [10]. При включении нейросетевых моделей в обработку текстов возрастает роль формализованного описания входных данных и критериев качества индексирования, поскольку итоговая разметка ключевыми терминами используется в последующих процедурах классификации, тематического анализа и навигации по корпусу научных публикаций [6]. Эмбеддинги в данной системе выступают элементом алгоритмической инфраструктуры, обеспечивающим сопоставление семантики терминов со структурными элементами научного текста [5]. На прикладном уровне автоматическое индексирование понимается как получение списка ключевых терминов, отражающих предметное содержание документа и поддерживающих сопоставимость материалов внутри коллекции при единых правилах обработки [9]. В задачах выделения ключевых терминов контекстные эмбеддинги получили распространение потому, что связь терма с окружением повышает корректность интерпретации полисемии и терминологических вариаций в научных публикациях [11]. Сравнение документов и подкорпусов переводит индексирование на уровень цифровой научной коллекции как совокупности взаимосвязанных материалов [7].

Кроме того, наличие в корпусах разнородных по форме описаний понятий, включая графовые и иерархические структуры, усиливает значимость согласования терминов с их функциональной ролью в документе [8]. В многоязычных массивах публикаций возрастает значение сопоставления обозначений понятий и именованных сущностей, поскольку совпадение смысла не гарантируется совпадением формы записи, а семантическое выравнивание напрямую влияет на качество индексирования [9]. Автоматическое индексирование в системном анализе рассматривается как последовательность процедур, сочетающих представление текста эмбеддингами, правила отбора кандидатов и критерии выбора финального набора ключевых терминов [1-2]. Перенос эмбеддингов между предметными областями приводит к расхождению распределений и смысловых соответствий, что снижает согласованность ключевых терминов в пределах одной коллекции [4]. Проверка методов на прикладных данных показывает значимость выбора процедуры группирования и параметров представления, поскольку близость векторов зависит от предобработки и применяемой модели [3]. Для задач коллекционного уровня ориентиром служат результаты кластеризации текстов с использованием эмбеддингов больших языковых моделей, что отражает влияние типа эмбеддингов на выделение тематически однородных массивов [13].

Материалы и методы

На первом этапе исследования происходит обработка полученных исторических данных экспериментального карьера. После этого основной набор данных разделяется на наборы обучения, тестирования и проверки. Затем набор данных подвергается четырем алгоритмам машинного обучения (SVM, ANN, DT и RF) с использованием инструмента Rapid Miner. Для создания и оценки прогноза используются четыре показателя: корреляция, абсолютная ошибка, относительная ошибка и среднеквадратическое значение. Процесс настройки параметров происходит на протяжении всего исследования, чтобы обеспечить наилучшую продуктивность четырех моделей.

Исходные данные об энергопотреблении были собраны на экспериментальном карьере. Различные измерения, включая напряжение, частоту, ток, коэффициент мощности, записывались, контролировалось их динамическое изменение. Каждую секунду устанавливалось новое рекордное значение энергопотребления. Данные были собраны с 21 участка экспериментального карьера на протяжении 1 суток, всего было сделано 3008 тыс. записей. После применения программы Python для фильтрации данных осталось 26,4 тыс. записей.

Результаты и их обсуждение

Автоматическое индексирование научных текстов опирается на представление ключевых терминов как компактных маркеров содержания, пригодных для сопоставления документов в цифровых научных коллекциях. Для этой цели векторные модели заменили ранние частотные схемы представления текста, поскольку числовое кодирование слов и фраз позволяет учитывать семантическую близость терминов, а не ограничиваться их встречаемостью [10]. Концептуально важным является разграничение дискретных представлений текста и непрерывных эмбеддингов, так как в задачах индексирования в отличие от простого подсчёта слов требуется приближение к смысловому ядру документа, выраженному ключевыми терминами [12]. Развитие контекстных моделей усилило роль контекста при интерпретации терминов, особенно в научных текстах с высокой терминологической плотностью и полисемией [9].

Уточнение понятия ключевых терминов в современных методах индексирования связано с тем, что качество разметки зависит от того, какая единица текста получает векторное представление и по каким правилам формируется итоговый набор терминов. В этом поле практическую значимость приобрели основанные на использовании эмбеддингов методы выделения ключевых терминов, ориентированные на сопоставление близости фраз к смысловому центру текста [11].

Для теоретического обоснования места векторных моделей важна связь «эмбеддинги — выбор модели — воспроизводимость результата», поскольку разные семейства эмбеддингов предоставляют различающиеся пространства близости и различающиеся списки ключевых терминов при одном и том же корпусе работ [5]. Как следствие, векторные модели в современных методах индексирования рассматриваются как техническое средство смыслового сопоставления терминов, задающее рамки для последующих алгоритмических решений [6].

В задачах автоматического индексирования на основе извлечения (extractive) формирование набора ключевых терминов сводится к отбору кандидатов из текста научной публикации с последующим ранжированием по заданному критерию близости к содержанию документа.

В качестве информационной базы для ранжирования обычно применяют представления текста, полученные при векторизации, тогда как сходство кандидатов с текстом оценивают в пространстве эмбеддингов, что удобно для системного анализа за счёт прозрачной формализации входных данных и метрики сопоставления [10]. Для подходов на основе извлечения характерна зависимость результата от качества сегментации, выбора кандидатов и от того, какие элементы научного текста попадают в обработку, поэтому при сопоставлении методов целесообразно рассматривать цепочку обработки как систему взаимосвязанных процедур с контролируемыми параметрами [12].

Подходы на основе порождения (generative) ориентированы на получение ключевых терминов как результата генерации, поэтому в итоговый набор могут попадать отсутствующие в тексте, но близкие по смыслу к тематике публикации формулировки.

Для системного анализа различие с подходами извлечения выражается в смене точки контроля, поскольку в подходах порождения основной риск связан с корректностью порождённых терминов и их соответствием исходному тексту, в то время как проверка качества опирается на сопоставление с авторскими ключевыми словами и семантическими метриками сходства [11].

В процессе использования контекстных векторных представлений оценка релевантности ключевых терминов опирается на контекст, что повышает точность интерпретации полисемии, однако возрастает значение процедур проверки и ограничений на генерацию [12].

В таблице ниже даётся сравнительная оценка подходов к автоматическому индексированию на основе извлечения и порождения (табл. 1).

В эмбеддинг-ориентированных методах индексирования ключевые термины оцениваются по близости векторных представлений, при этом часто возникает эффект равноправия векторов, когда единицы текста получают сопоставимый «вес» независимо от роли в научной публикации [12]. Контекстная размытость проявляется в том, что близость векторов отражает общий тематический фон фрагмента, а граница между термином как смысловым ядром и термином как частным упоминанием различается слабо, вследствие чего в итоговый список попадают второстепенные обозначения [11]. Утрата структурной значимости терминов дополнительно связана с тем, что одинаковая лексема в аннотации, заголовке или разделе результатов получает близкие векторы, хотя для индексирования научных текстов значение таких позиций различается [10]. Для смягчения указанных ограничений в индексирование вводится взвешивание

Таблица 1.

Сравнительная характеристика методов извлечения и порождения в автоматическом индексировании научных текстов

Параметр	Извлечение	Порождение	Системный анализ
Выходной результат	Список ключевых терминов из текста	Список ключевых терминов, включая отсутствующие	Различие объекта контроля качества
Основание отбора	Кандидаты из текста	Генерация последовательности терминов	Различие источника множества кандидатов
Источник ошибок	Потеря релевантных кандидатов на этапе выделения	Появление терминов вне содержания публикации	Различие доминирующего риска
Прозрачность процедуры	Высокая интерпретируемость шагов отбора	Снижение интерпретируемости внутренних представлений	Различие управляемости параметров
Критерий ранжирования	Метрика близости кандидата и текста	Метрика качества генерации и близости к тексту	Различие схемы оценивания
Зависимость от структуры текста	Сильная зависимость от сегментации и границ фрагментов	Зависимость от представления документа в модели	Различие чувствительности к препроцессингу
Контроль терминологии	Контроль словаря за счёт ограничений на кандидатов	Контроль словаря через ограничения генерации	Различие инструмента нормализации
Метка качества	Сходство с авторскими ключевыми словами; точность отбора	Сходство с авторскими ключевыми словами; семантика	Сопоставимость метрик при едином наборе данных
Применимость к коротким фрагментам	Снижение качества при малом числе кандидатов	Повышение качества при наличии сильных контекстных представлений	Различие зависимости от объёма текста
Воспроизводимость результата	Высокая при фиксированных правилах отбора	Зависимость от стохастики генерации и параметров	Различие воспроизводимости в экспериментах
Масштабирование на коллекцию	Стабильное при унификации правил извлечения	Чувствительность к доменной вариативности коллекции	Различие рисков при росте неоднородности данных

Источник: авторская разработка

векторного представления терма с учётом его повторов и структурной позиции в тексте, что позволяет отделять терминологическое ядро от фоновых употреблений и учитывать вклад отдельных фрагментов [13]. Взвешенное векторное представление терма в общем виде можно записать формулой (1)

$$v^w(t) = \frac{\sum_{i=1}^{n} \alpha_i \times v_i(t)}{\sum_{i=1}^{n} \alpha_i} \quad (1)$$

где: $v_i(t)$ — векторное представление терма t в i-м контексте; α_i — весовой коэффициент, отражающий значимость позиции терма в структуре научного текста; n — число контекстных употреблений терма в документе.

Особенности векторного представления терминов формируют систему алгоритмических ограничений и условий применимости эмбеддинг-методов автоматического индексирования (табл. 2).

Структурно-семантическая модель автоматического индексирования научных текстов на основе эмбеддингов предполагает раздельную обработку смысловых единиц текста и элементов структуры документа, поскольку для ключевых терминов значимость определяется не только лексическим содержанием, но и позицией в публикации [12]. Прикладная реализация опирается на поэтапное представление документа, при котором векторные представления вычисляются для фрагментов, связанных со структурой научной публикации, а затем сводятся к единому описанию документа для последующего отбора ключевых терминов [10]. При работе с цифровыми научными коллекциями значимым становится согласование уровня фрагмента и уровня документа, поскольку именно на уровне документа выполняется сопоставление материалов коллекции при поиске и тематической группировке [13]. Агрегация векторных представлений фрагментов в агрегированный вектор документа задаёт-

Таблица 2.

Алгоритмические ограничения и условия применимости эмбеддинг-методов автоматического индексирования научных текстов

Ограничение	Содержание проблемы	Практическое следствие	Условие применимости
Равноправие векторов	Сходные векторы для терминов разной значимости	Попадание второстепенных обозначений в ключевые термины	Введение весовых коэффициентов для позиций текста
Контекстная размытость	Смешение тематического фона и смыслового ядра	Снижение точности отбора терминов	Учёт локального и глобального контекста
Утрата структурной значимости	Игнорирование роли заголовков и разделов	Потеря приоритетов терминов документа	Моделирование структуры научного текста
Зависимость от домена	Смещение смыслов при переносе моделей	Нестабильность результатов между коллекциями	Доменная адаптация векторных моделей
Чувствительность к предобработке	Влияние токенизации и нормализации	Различия в списках ключевых терминов	Унификация процедур подготовки данных
Масштабность корпуса	Рост вычислительной сложности	Замедление индексирования больших коллекций	Оптимизация и выбор облегчённых моделей
Полисемия терминов	Смешение разных значений одной формы	Ошибки семантического сопоставления	Применение контекстных представлений
Неполнота множества кандидатов	Пропуск значимых формулировок	Ограничение точности извлечения	Расширенные правила генерации кандидатов
Стабильность результата	Вариативность при смене параметров	Снижение воспроизводимости	Жёсткая фиксация настроек эксперимента
Интерпре-тируемость	Трудность объяснения выбора терминов	Ограничение экспертного контроля	Комбинация векторов и правил отбора
Шум текстовых данных	Наличие нетерминологических элементов	Засорение итогового списка	Фильтрация и тематическая нормализация
Многоязычность коллекций	Несоответствие форм терминов	Потеря сопоставимости документов	Семантическое выравнивание языков

Источник: авторская разработка

ся взвешенной комбинацией с учётом структуры текста, что обеспечивает разделение вкладов заголовка, аннотации и основного текста в итоговом представлении [12]. Формула (2) представлена следующим выражением

$$v(d) = \frac{\sum_{s=1}^{m} \beta_s \times v_s(d)}{\sum_{s=1}^{m} \beta_s} \quad (2)$$

где: вектор $v_s(d)$ соответствует представлению фрагмента документа, индекс s относится к структурному элементу, коэффициент $ß_s$ задаёт вес структурного элемента, m соответствует числу учитываемых элементов структуры научного текста.

На основании рассмотренных вводных целесообразно разработать структурно-семантическую модель автоматического индексирования научного текста на основе использования эмбеддингов (рис. 1).

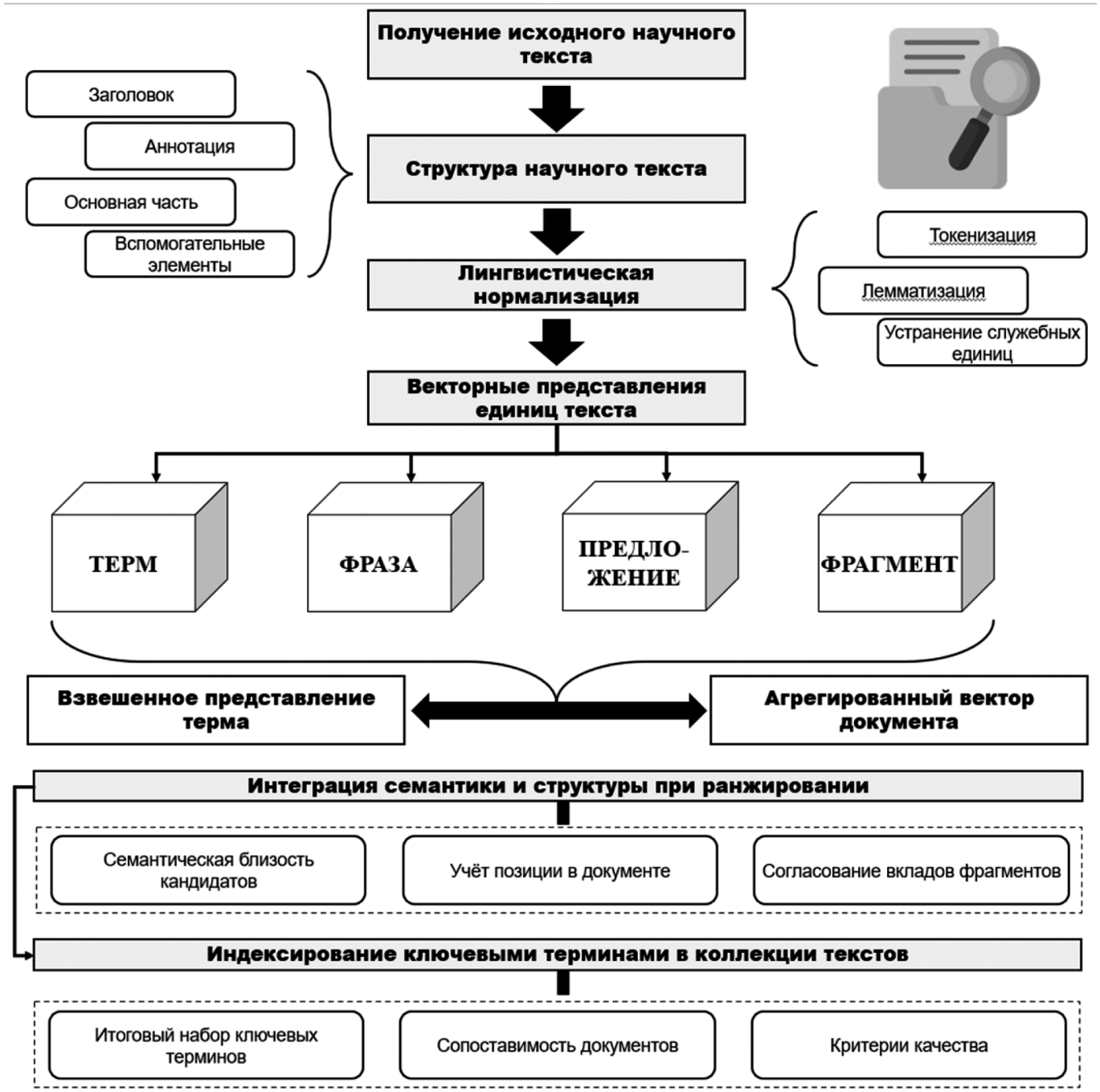

Рис. 1. Структурно-семантическая модель автоматического индексирования научного текста на основе векторных представлений

Источник: авторская разработка

Представленная авторская схема отражает целостную структурно-семантическую модель автоматического индексирования научных текстов на основе эмбеддингов, в которой процесс обработки организован как взаимосвязанная система уровней и блоков. Исходный научный текст рассматривается в неразрывной связи с его внутренней структурой, после чего выполняется лингвистическая нормализация, обеспечивающая единообразие представления единиц текста. Далее формируются векторные представления термов, фраз, предложений и фрагментов, что позволяет перейти от текстовой формы к формализованному описанию содержания. На следующем уровне осуществляется взвешивание представлений термов и построение агрегированного вектора документа с учётом значимости различных структурных элементов публикации. Интеграция семантики и структуры при ранжировании кандидатов обеспечивает согласование близости терминов с их позицией в документе и с вкладом отдельных фрагментов. Финальный этап связан с формированием итогового набора ключевых терминов, поддерживающих сопоставимость документов и применение единых критериев качества в рамках цифровых научных коллекций.

Качество автоматического индексирования научных текстов зависит от того, какие структурные элементы документа включаются в обработку и с какими весами выполняется агрегация векторных представлений. При равном учёте заголовка, аннотации и основной части возрастает вероятность смещения ключевых терминов в сторону наиболее частотных или тематически «общих» формулировок, тогда как для научной публикации принципиальна связь термина с информативными фрагментами, отражающими предметный вклад работы [12]. В процессе раздельной обработки структурных элементов и последующей интеграции оценка релевантности кандидатов опирается на более строгий контекст, поэтому снижается доля терминов, совпадающих по тематике, однако не отражающих содержание конкретного исследования [10].

Наибольшая концентрация смысловых маркеров обычно наблюдается в заголовке и аннотации, однако такая концентрация приводит к появлению «витринных» терминов, ориентированных на описание области, а не на детализацию результата, поэтому приоритет этих элементов оправдан только при контроле структуры и при сопоставлении с основной частью [11]. Секции постановки задачи и обзора нередко воспроизводят общепринятую терминологию предметной области, что повышает риск доминирования терминов высокого уровня и вытеснения более специфичных обозначений, встречающихся в разделах методики и результатов [9].

Проверка воспроизводимости результатов в рамках цифровых научных коллекций опирается на согласованность ключевых терминов при варьировании долей структурных элементов и при смене процедур нормализации, поэтому для сопоставимых экспериментов требуется единый протокол подготовки корпуса и единые правила формирования агрегированного вектора документа [13]. При нарушении сопоставимости входных условий расхождения в пространстве векторных представлений приводят к неустранимым различиям в ранжировании кандидатов, что усложняет оценку качества автоматического индексирования как процедуры системного анализа [12].

Выводы

В исследовании рассмотрено автоматическое индексирование научных текстов ключевыми терминами как задача системного анализа, в которой важны сопоставимость результатов и единые правила обработки цифровых научных коллекций. Разработанная структурно-семантическая модель задаёт единое представление документа, в котором семантическая близость кандидатов согласуется с их позицией в публикации. Практическая ценность предложенных решений связана с возможностью применения чётких критериев проектирования процедур автоматического индексирования, ориентированных на воспроизводимость и интерпретируемость результатов в цифровых научных коллекциях. Для прикладного применения важна стандартизация подготовки текста и правил агрегации, поскольку сопоставимость итоговых наборов ключевых терминов зависит от единообразия входных процедур и выбранных весов структурных элементов.

ЛИТЕРАТУРА

1. Гусев Д.И., Апанович З.В. Влияние методов построения векторных представлений имен сущностей на качество выравнивания сущностей // Научный сервис в сети Интернет. — 2022. — Т. 24. — С. 155–166.
2. Гусев Д.И., Апанович З.В. Как эмбеддинги имен сущностей влияют на качество выравнивания сущностей // Электронные библиотеки. — 2023. — Т. 26. — № 1. — С. 52–79.
3. Демидова Л.А., Морошкин Н.А. Решение задачи кластеризации векторных представлений регулярных выражений // Вестник Воронежского государственного технического университета. — 2025. — Т. 21. — № 2. — С. 50–59.
4. Комарова Л.А., Колосов А.М., Соловьев В.И. Сопоставление векторных представлений вакансий и резюме с использованием больших языковых моделей // International Journal of Open Information Technologies. — 2025. — Т. 13. — № 2. — С. 56–66.
5. Куровский С.В., Мишин Д.А., Маринин А.К., Бурдик В., Куровская М.А. Современные подходы к автоматизации и оптимизации инвестиционных сервисов телекоммуникационных компаний // Инновации и инвестиции. — 2024. — № 10. — С. 455–460.

6. Куровский С.В., Мишин Д.А., Штыков Р.А. Задачи и методы формализации и оптимального управления цифровыми сервисами в компаниях // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. — 2024. — № 10–2. — С. 39–45.
7. Куровский С.В., Мишин Д.А., Шульман В.Д. Алгоритм балансировки нагрузки в гетерогенной среде вычислительной системы // Международный научно-исследовательский журнал. — 2025. — № 5 (155). — С. 1–10.
8. Лях А.П. Классификация и основные алгоритмы эмбеддинга в контексте больших языковых моделей // Ученые заметки ТОГУ. — 2024. — Т. 15. — № 3. — С. 79–83.
9. Прошина М.В. Современные методы обработки естественного языка: нейронные сети // Экономика строительства. — 2022. — № 5. — С. 27–42.
10. Усатов А.А., Недзьведь А.М., Цзижань Г. Оценка сходства между наборами данных с помощью векторных представлений // Доклады Белорусского государственного университета информатики и радиоэлектроники. — 2025. — Т. 23. — № 3. — С. 70–76.
11. Чернышева А.В., Хлопотов М.В. Метод векторного представления и автоматизированной оценки учебного плана в системе интеллектуальной поддержки проектирования образовательных программ // Экономика. Право. Инновации. — 2024. — №. 2. — С. 61–71.
12. Asudani D.S., Nagwani N.K., Singh P. Impact of word embedding models on text analytics in deep learning environment: a review // Artificial intelligence review. — 2023. — Vol. 56. — No. 9. — P. 10345–10425.
13. Egger R. Text representations and word embeddings: Vectorizing textual data // Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications. — Cham: Springer International Publishing, 2022. — P. 335–361.
14. Khan M.Q. et al. Impact analysis of keyword extraction using contextual word embedding // PeerJ Computer Science. — 2022. — Vol. 8. — P. 1–16.
15. Patil R. et al. A survey of text representation and embedding techniques in nlp // IEEe Access. — 2023. — Vol. 11. — P. 36120–36146.
16. Petukhova A., Matos-Carvalho J.P., Fachada N. Text clustering with large language model embeddings // International Journal of Cognitive Computing in Engineering. — 2025. — Vol. 6. — P. 100–108.

© Комаров Иван Дмитриевич (i.komaroni@ya.ru)
Журнал «Современная наука: актуальные проблемы теории и практики»